

Rubric-Guided Generative AI for Scalable Creativity Assessment in Educational Games

Seyedahmad Rahimi, CLASS Lab, College of Lifetime Learning, Georgia Tech., srahimi@gatech.edu

Hongming Li, School of Teaching and Learning, University of Florida, hli3@ufl.edu

Salah Esmailigoujar, School of Teaching and Learning, University of Florida, s.esmailigoujar@ufl.edu

Deniz Ercan, School of Teaching and Learning, University of Florida, deniz.ercan@ufl.edu

Abstract: Assessing creativity at scale remains challenging in learning sciences. This study examined whether GPT-4o could assess creativity in 421 student-designed levels from *Physics Playground*. Using a structured rubric, GPT-4o ratings were compared with human raters. Results showed strong correlations ($r = .74$) with multimodal inputs, and reliability analyses showed consistent performance across runs. Findings suggest that rubric-guided LLMs can approximate human creativity judgments efficiently. This approach offers a promising method for scalable automated creativity assessment.

Introduction

Creativity is a cornerstone of human progress and a key competency for thriving in the twenty-first century (OECD, 2024). Organizations such as UNESCO and the OECD emphasize its role in innovation and navigating complex, technology-rich societies (UNESCO, 2023; OECD, 2024). In education, creativity supports forming original connections and deepening understanding, especially in STEM (Luchini et al., 2025). It is commonly defined as producing ideas or artifacts that are both novel and appropriate—original, functional, and valuable (Amabile, 2018; Kaufman & Beghetto, 2009; Plucker et al., 2004). As artificial intelligence (AI) evolves, it reshapes how creativity is expressed, generated, and evaluated (Rahimi et al., 2026). Generative AI (GenAI) tools powered by large language models (LLMs) can produce text, images, and music that approximate human originality (Beaty et al., 2023; Heigl, 2025) and expand idea generation in learning environments (Cai & Gao, 2025). At the same time, their growing use raises questions about distinguishing human and AI-generated creativity and evaluating creativity in AI-supported contexts (Beghetto & Karwowski, 2023; Rahimi et al., 2026). Although AI lacks lived experience and emotion (Lockhart, 2025; Runco, 2023), it can function as a collaborative partner that augments human creativity (Haase & Pokutta, 2024). In educational settings, GenAI can support brainstorming, visualization, and prototyping, but it also increases the need for scalable and ecologically valid creativity assessment. Traditional measures such as the Torrance Tests of Creative Thinking (TTCT; Torrance, 1996) are labor-intensive and poorly suited to dynamic environments (Rafner et al., 2023). AI-based approaches, including rubric-guided LLM evaluation, offer a way to assess creativity unobtrusively within authentic contexts such as digital games (Rafner et al., 2023; Rahimi, 2025).

This study investigates whether an LLM can accurately assess the creativity of over four hundred human-designed game levels in *Physics Playground*. Using rubric-guided prompting and multimodal inputs, we examine the validity and reliability of GPT-4o's creativity scores compared to expert human ratings with the following research question: *To what extent GPT-4o can produce valid and reliable assessments of human creativity in the context of the Physics Playground level editor?*

Related work

Creativity assessment has traditionally relied on standardized instruments such as the TTCT and the Consensual Assessment Technique (CAT; Amabile, 1982). Although theoretically grounded, these methods have notable drawbacks. The TTCT and similar tests often decontextualize creativity and focus on verbal or figural divergent thinking rather than authentic, domain-specific performance (Barbot et al., 2016). The CAT, while valid, requires multiple expert raters which makes it costly and impractical for large-scale use (Kaufman et al., 2008). LLMs such as GPT-4 offer new potential for creativity assessment due to their ability to interpret complex multimodal inputs and generate human-aligned evaluations (Beaty et al., 2023). Among these, prompt engineering is most cost-effective for educational contexts. Yet three critical gaps remain: Prior work focuses on text-based products rather than multimodal artifacts; uses small datasets rather than large-scale authentic data; and has not examined single versus multiple evaluation reliability. This study addresses these gaps using *Physics Playground*'s level editor.

Method

This study examined whether GPT-4o can assess human creativity in 421 game levels created in *Physics Playground* (Shute et al., 2019), a 2D educational game where players design physics-based machines (e.g., ramps, pendulums) to guide a ball to a balloon. Using the built-in level editor, participants created original levels in an open-ended environment (Rahimi, 2020; Rahimi & Shute, 2021). The dataset consisted of 421 levels designed by college students from a prior study (Rahimi, 2020). After a brief tutorial, participants had up to two hours to create playable levels and then selected their most creative designs. Each level was saved as both an image and a JSON file containing full structural data.

Creativity was evaluated using a seven-item rubric adapted from Rahimi and Shute (2021), covering originality, elaboration of lines and objects, aesthetics, humor/surprise, and title creativity. Two educational experts independently rated each level on a 0–2 scale per dimension (maximum total = 13). Final human scores were computed by averaging the two raters' scores for each dimension, serving as the ground truth. GPT-4o was prompted to evaluate each level using the same rubric and to generate structured JSON outputs with dimension-level and total scores. Three prompting conditions were tested, differing only in input modality while keeping rubric text and output format constant. Validity was assessed using Spearman's rank-order correlation (ρ) between model and human scores, with confidence intervals obtained via bootstrapping (1,000 resamples). Reliability was examined by comparing a single model run to the average of three independent runs.

Results

Results showed that rubric-guided prompting substantially improved the accuracy of the model's assessments (see Table 1). When GPT-4o used the validated creativity rubric, its total creativity scores showed a strong positive correlation with human ratings ($r = .74$), while scores generated without the rubric displayed only moderate agreement ($r = .61$). Further analyses revealed that including both the level image and its structured JSON data produced the strongest correspondence with human evaluations. Removing the image reduced the model's ability to detect nuanced creative details such as aesthetics or humor, which suggests that visual information played a key role in capturing human-like judgment.

Table 1

Correlation (Spearman's ρ) of Prompting Methods with Human Ratings on the Dataset ($N = 421$)

Dimension	M1 vs Human	M2 vs Human	M3 vs Human
Solvability	.24 [-.01, .66]	.35 [-.01, .82]	-.01 [-.01, -.00]†
Line Elaboration	.64 [.54, .73]	.61 [.50, .70]	.56 [.45, .66]
Object Elaboration	.73 [.68, .78]	.71 [.66, .75]	.67 [.61, .72]
Aesthetics	.64 [.57, .70]	.58 [.51, .64]	.62 [.56, .68]
Humor & Surprise	.57 [.50, .63]	.45 [.37, .52]	.47 [.39, .54]
Title Creativity	.35 [.26, .44]	.34 [.25, .42]	.34 [.26, .43]
Originality	.53 [.46, .60]	.52 [.44, .59]	.48 [.40, .55]
Total Score	.74 [.69, .79]	.71 [.65, .768]	.68 [.62, .74]

Note: All correlations are significant at $p < .001$ (FDR-corrected), unless marked with † ($p > .05$). Values in brackets are 95% confidence intervals estimated via bootstrapping (1000 resamples).

Reliability analyses indicated that a single call to GPT-4o evaluation produced results highly consistent with those obtained from multiple independent runs ($\rho = .74$ vs .71; RMSE = 1.9 vs 2.1). Comparisons among related models confirmed that larger architectures performed better: GPT-4o > GPT-4o-mini > GPT-4.1-nano, with overall correlations of approximately .71, .62, and .41, respectively. Overall, these results demonstrate that with a clear rubric and multimodal input, an LLM can approximate expert evaluation of creative quality across hundreds of human-generated game levels.

Discussion and conclusion

The findings show that LLMs can provide valid and reliable assessments of creativity when guided by structured rubrics and multimodal input in the context of *Physics Playground*. GPT-4o's strong correlation with expert ratings suggests that it can recognize creative qualities such as originality, elaboration, and aesthetic appeal across a large number of student-designed game levels. These results also highlight the importance of visual context in creative assessment. The model's performance declined when images were excluded. This result implies that creativity in visual-spatial artifacts cannot be captured through structured data alone. This study demonstrates that LLMs, when guided by a validated rubric, can generate creativity ratings that closely align with human judgments in a non-asset-based educational game environment. These results underscore the effectiveness of rubric-guided prompting in ensuring validity and reliability, highlighting the potential of LLMs as scalable, unobtrusive assessors of creativity.

References

- Amabile, T. M. (1982). *Social psychology of creativity: A consensual assessment technique*. *Journal of Personality and Social Psychology*, 43(5), 997–1013. <https://doi.org/10.1037/0022-3514.43.5.997>
- Amabile, T. M. (2018). *Creativity in context: Update to the social psychology of creativity*. Routledge.
- Barbot, B., Lubart, T., & Besançon, M. (2016). *Creative potential in educational settings: Its nature, measure, and nurture*. *Education* 3-13, 44(4), 427–435. <https://doi.org/10.1080/03004279.2015.1020643>
- Beaty, R. E., Kenett, Y. N., Hass, R. W., & Schacter, D. L. (2023). Semantic memory and creativity: The costs and benefits of semantic memory structure in generating original ideas. *Thinking & Reasoning*, 29(2), 305–339. <https://doi.org/10.1080/13546783.2022.2076742>
- Beghetto, R. A., & Karwowski, M. (2023). *Creative self-beliefs: From creative potential to creative action*. In R. J. Sternberg & J. C. Kaufman (Eds.), *Handbook of organizational creativity* (2nd ed., pp. 179–193). Academic Press. <https://doi.org/10.1016/B978-0-323-91840-4.00010-4>
- Heigl, R. (2025). Generative artificial intelligence in creative contexts: a systematic review and future research agenda. *Management Review Quarterly*, 1–38. <https://doi.org/10.1007/s11301-025-00494-9>
- Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). *A comparison of expert and nonexpert raters using the consensual assessment technique*. *Creativity Research Journal*, 20(2), 171–178.
- Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of general psychology*, 13(1), 1–12. <https://doi.org/10.1037/a0013688>
- Luchini, S. A., Maliakkal, N. T., DiStefano, P. V., Laverghetta Jr, A., Patterson, J. D., Beaty, R. E., & Reiter-Palmon, R. (2025). Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*.
- OECD (2024). *OECD Economic Outlook, Volume 2024 Issue 2*, OECD Publishing, Paris.
- Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational psychologist*, 39(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1
- Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836–1838. <https://doi.org/10.1038/s41562-023-01751-1>
- Rahimi, S. (2020). *Inspire, instruct, or both? Game-based support of creativity* [Doctoral dissertation, Florida State University].
- Rahimi, S. (2025). Going Beyond the Brick: Assessing and Supporting Creativity Using AI-Powered Digital Games. *Creativity Research Journal*, 37(2), 275–283. <https://doi.org/10.1080/10400419.2023.2241779>
- Rahimi, S., Dede, C., Esmailigoujar, S., & Babae, M. (2026). Augmenting human creativity with responsible and ethical use of generative AI. In M. J. Worwood & J. C. Kaufman (Eds.), *Generative artificial intelligence and creativity* (Explorations in Creativity Research, pp. 87–99). Academic Press. <https://doi.org/10.1016/B978-0-443-34073-4.00010-1>
- Rahimi, S., & Shute, V. J. (2021). First inspire, then instruct to improve students' creativity. *Computers & Education*, 174, 104312. <https://doi.org/10.1016/j.compedu.2021.104312>
- Shute, V. J., Almond, R. G., & Rahimi, S. (2019). *Physics Playground* (Version 1.3) [Computer software]. Tallahassee, FL: Retrieved from <https://pluto.coe.fsu.edu/ppteam/pp-links/>
- Torrance, E. P. (1966). Torrance tests of creative thinking. *Educational and psychological measurement*.
- UNESCO. (2023). *Global education monitoring report 2023: Technology in education—A tool on whose terms?* UNESCO. <https://www.unesco.org/gem-report/en/publication/technology>